LOH Detection on Allele Ratio in Partek® Genomics Suite™

You can detect LOH from allele ratio spreadsheet by choosing **Tools>Detect LOH**

Allele Ratio observations

Allele Ratio values are given between 0 and 1. In regions of heterozygousity, we expect to observe a distribution with modes near 0, .5, and 1, driven by AA, AB, BB calls. In regions with allele imbalances, we expect to observe four modes, symmetric around .5.

Likelihood of Allele Ratio observations

The probability of each observed allele ratio value, x, can be divided into three terms.

$P(x) = P(x \mid het)\, P(het) + P(x \mid hom)\, P(hom) + P(x \mid err)\, P(err)$

P(het), P(hom), P(err) can be viewed as mixture parameters defining the relative balance of heterozygous, homozygous, and errors in the data.

P(x | het) is the likelihood of observing x, given that it is an observation at a heterozygous SNP. The chance of observing a high or low value from x are considered equally likely (A and B are arbitrary labels). So, x is assumed to be drawn from one of two distributions (symmetric around .5) when we have a heterozygous mean.

x ~ Normal(mean, variance)

OR

x ~ Normal(1 – mean, variance)

Identical models are applied to the homozygous likelihood using the homozygous mean with the same variance. This may not be an appropriate assumption for data points clamped at 0 and 1, however, it does perform reasonably in practice.

Since x is over the interval 0 to 1 and we expect uniform errors,

$P(x \mid err) = 1$

Analyzing windows of Allele Ratio observations

Now that we have determined the likelihood of each allele ratio observation, we now examine overlapping windows of allele ratio estimates at consecutive SNPs. The window size is a user specified parameter. We use an EM algorithm to maximize the likelihood of observing the data in each window. The parameters estimated are heterozygous mean (range 0 - .5), homozygous mean (range 0 -

.5), variance, and heterozygousity/homozygousity mixture parameters. The mean and variance estimates define four Gaussian modes blended using the mixture parameters. The error mixture parameter, P(err), is fixed at .01.

Each window is assigned a score using the heterozygous and homozygous means as,

Score = .5 * (mean_het – mean_hom) / (.5 – mean_hom)

The score can be interpreted as values near 0 implying LOH, and values near .5 representing retention of heterozygousity.

Segmenting Window Scores

After generating scores for all windows on a chromosome, a bottom up segmentation algorithm is applied by seeding each point to belong to its own region of one marker with sum squared error of 0. Neighboring regions are combined in an order that would introduce the smallest increase in sum squared error until all remaining join candidate would introduce more error than an algorithm threshold. This threshold is set as the window size of 10 markers.

To report region boundaries, we consider each window genomic coordinates to be its mid-point. The boundary between two adjacent windows that belong to different regions is reported as the point between the two windows' position.